Early warning of pirate attacks based on decision tree[1]

Sun Mao-jin[1,2], Lv Jing[1], Gao Tian-hang[1], Sun Xiao-shan[1]

Abstract. Considering the pirates threats to the international transport ships, a constructed forewarning model of pirates attack can provide warning information before it happens, and it will reduce the probability of attack substantially. Therefore, a forewarning model based on decision tree is constructed to study the correlation between the severity of pirates attack and other factors. This method takes the ships types and geographic areas as input data, pirates attack consequence as output data, and the pirates attack in the waters of east Africa as study subjects. After the validation, the results show that the predictions are much better than random guesses. This proves that the method is effective and scientific.

Key words. Ship security, pirates attack, forewarning, decision tree.

1. Introduction

As the trades among various countries in the world gets frequent, marine transport has been becoming more and more important in international society, which has turned into one of the foremost transportation modes. However, transport ships have been affected greatly by the traditional and unconventional security threats. The security issue is one of the problems the practitioners concern most. With the development of technologies, the security issues of marine transport system, primarily caused by climate and geographical factors, have been decreased greatly. But the unconventional security threats still exist violently, which is mainly reflected as piracy and maritime terrorism. By taking the piracy incidents with most extensive impact as examples, as for the international waters, the piracy activities frequently occurs in several areas, including Aden Gulf, Indian Ocean, West Africa coast, Malacca and sea areas belonging to South China Sea. To deal with these situations, it is particularly necessary to make clear how to build the piracy forewarning

[1] College of Transportation Management, Dalian Maritime University, Dalian 116026, Liaoning, China

[2] Corresponding author

model by reasonable algorithm for warning the passing transport ships before pirates attack based on the current information, guiding them to deviate and take effective measures.

At present, the direction of research on the pirate attacks are diversified, covering politics, society, military affairs and management, etc. For settling the matter of Somali pirates, [1] has discussed and analyzed its multifaceted problem, such as formation history, organization and operation, countermeasures taken by various countries internationally, etc. [2] has studies the economic losses of marine transport ships brought about by the pirate attacks. [3] treats the private attacks as one of dynamic behaviors, analyzes the number and movement area of privates by using the dynamic model, and takes the areas in Aden Gulf as case to study. Literature [4] proposes a new escort maneuver for warships combating pirates by adopting the idea about covering priority area first, so as to improve navy's combat efficiency. References [5—6] respectively uses the Stackelberg game and evolutionary game to simulates the game process between navy and pirates; afterwards, naval escort area has been selected while the decision-making process between both of them has been refined. Reference [7] adopts the game theory to study the fairway programming in order to provide the navigation route, of which can avoid the private attacks, for the transport ships. There are no researches on forewarning of privates attacks despite the fact that privates attacks studies are relatively plentiful. To build a better forewarning model of pirate attacks, this paper chooses the researches related to forewarning as references. Paper [8] applied the combination of fuzzy comprehensive evaluation and analytic hierarchy process (AHP) to forewarning classification of public emergencies; on that basis, forewarning classification model is formed. Reference [9] built the BP neural network model by using neural network method based on the characteristics of petroleum safety precaution. Paper [10] built the alarm index system and forewarning model for spatial agglomeration suitability of manufacturing industry by catastrophe series method, which achieved great forewarning results. Reference [11] set up the incident-based forewarning system of supply chain so as to supervise and alert the timeliness of supply chain. Liu et al. [12] created a public information space for urban flood disaster in order to forewarn this kind of natural hazard. In conclusion, many attempts in the other forewarning study are done by the relevant scholars, which offers the formed thoughts and approaches that play an enlightening role in the research of this paper. This paper builds a forewarning model of pirate attacks by selecting the proper method and referring to the historical data characteristics and forming processes of pirate attacks

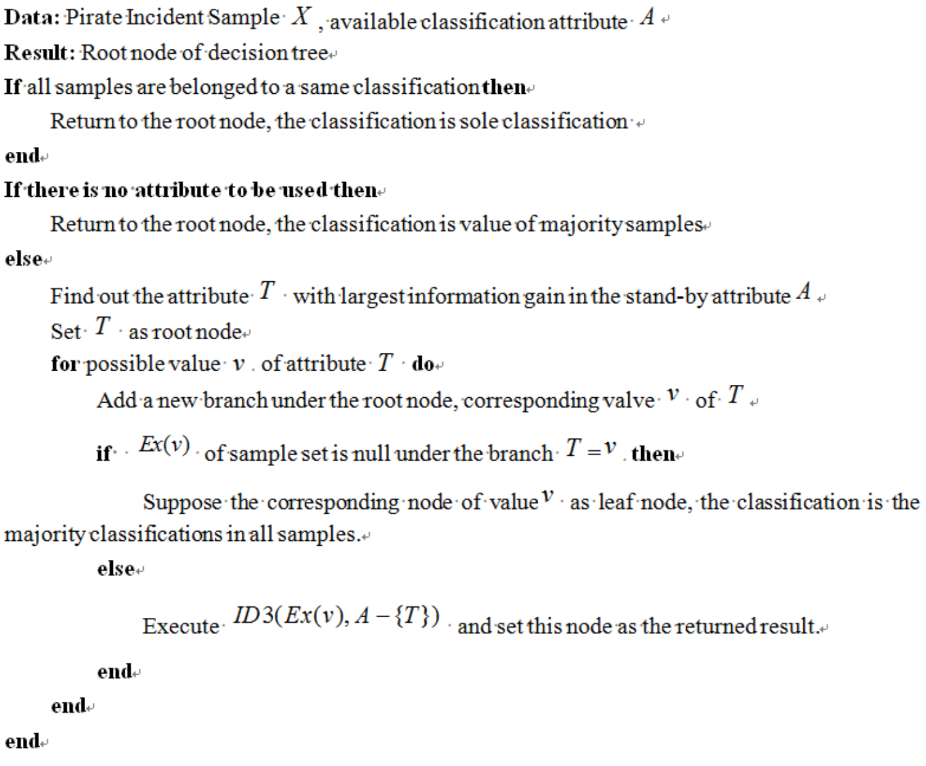
1.1. Decision tree overview

Decision tree is a common method with supervised classification in machine learning. In a generated decision tree, a node represents an attribute, and its branch is the optional value of corresponding attribute. When a sample is going to be classified, node will be sought out in accordance with the attribute value correspondingly based on the sample characteristics; and then, the previous step shall be run repeatedly till leaf node is reached. Prediction classification of this sample is the leaf classification, thereby the classification is completed. As for construction algorithm of decision tree, this paper set ID3 algorithm as example. ID3 is one of algorithms constructing the decision tree that are built on the information theory while taking the information entropy and information gain degree as valuation criterion. Intuitively, a higher information gain can distinguish various samples better. Sample points are classified to a node (e.g. root node); ID3 calculates the information gain partitioned by each attribute. Afterwards, the attribute with highest gain is selected as the classification attribute, which is also the standard to partition the sample to the next child node of corresponding value, and then this step should be repeated starting from the child node till there are no more attributes to be used, or new information gain to be partitioned.

Mathematical formulation of decision tree: the expression of information entropy is $H(S) = -\sum_{x\in X} p(x)\log_2 p(x)$. Symbol S represents the data set, X is the classification result of S, x is one member set of classifications in S, $p(x)$ is the proportion of Class x. It can thus be seen that, when there is only one classification in S, the information entropy is 0. But if there are more classification in S, the more the proportion, the higher the information entropy. That is to say, the information entropy reflects the impurity level of data. The information gain is built on the information entropy, thus it shows the change of information entropy both before and after data set is partitioned, namely the impurity level variation of pre- and post-data. The information gain is represented by $IG(A)$. In the formulation $IG(A) = H(S) - \sum_{t\in T} p(t)H(t)$, where $H(S)$ is the information entropy before partitioned, t is one of multiple data set get from the partitioned attribute A, T is aggregate of these date sets, $S = \bigcup_{t\in T} t$, $p(t)$ is the proportion of member numbers of t in S, and $H(t)$ is the information entropy of data set t. If a given data set is partitioned by different attributes, the information entropy results are always different. The attribution with higher information gain enables to get nodes with higher purity after partitioned. Actually, ID3 adopts this logic to lower the classified information entropy fast.

1.2. Forewarning model construction for pirate attacks

In the forewarning model of pirate attacks, the navigation risk is predicted and judged according to the characteristics of this navigation, such as ship types, the waters covered by the lane, longitude and latitude, the relation with coastline, etc. Longitude and latitude, a single data, cannot provide any valuable information. Hence, this paper selects the space region formed by integrated conditions as characteristics of navigation risk. Specifically, the entire waters are divided into multiple areas based on the pirate attack time, and each area contains a class of pirate attack incidents. One classification has various elements related to pirate attack possibly. The elements are similar with each other in one classification, but different among classifications. This classification represents the different characteristics of pirates in different areas, for example, some pirates, who value the money, are used to kidnap the hostages for getting ransom; and some others do not care about sailors' lives, they will kill the sailors who put up a desperate struggle when they are opposed. This situation is ascribed to the different movement range of pirates. Obviously, pirate characteristics in different areas have done a great deal to affect the danger levels in different areas. Except the basic characteristics, like ship types, the factor of pirate having different characteristics in different areas must be considered, too; in the meantime, ships sailing navigating into an area should be added into the judgment conditions of decision tree. Thus, for decision tree method, the best characteristic should be filtered for running the partition, which conforms to the conditions about one of attribute value and risk level. And then a new attribute will be searched to its child node. The decision tree is going to be constructed in this way. When steps are done, a navigation characteristic is put into the generated decision tree, finally getting a risk level as the prediction result. The specific algorithm of constructing a decision tree is shown as Fig. 1.



```
Data: Pirate Incident Sample $X$, available classification attribute $A$
Result: Root node of decision tree
If all samples are belonged to a same classification then
    Return to the root node, the classification is sole classification
end
If there is no attribute to be used then
    Return to the root node, the classification is value of majority samples
else
    Find out the attribute $T$ with largest information gain in the stand-by attribute $A$
    Set $T$ as root node
    for possible value $v$ of attribute $T$ do
        Add a new branch under the root node, corresponding valve $v$ of $T$
        if $Ex(v)$ of sample set is null under the branch $T=v$ then
            Suppose the corresponding node of value $v$ as leaf node, the classification is the
majority classifications in all samples.
        else
            Execute $ID3(Ex(v), A-\{T\})$ and set this node as the returned result.
        end
    end
end
```

Fig. 1. ID3 algorithm

2. Case study

To demonstrate that the forewarning model of pirate attack constructed based on the decision tree algorithm in this paper has certain practical value, this paper selects the statistic pirate attacks of GISIS database occurred in the waters of East Africa as research objects, and set the ship types and geographic areas as the forewarning information for verifying the pirate-attack-result-oriented forewarning result.

2.1. Data processing

The first step is to quantify the data. The quantitative method is determined by the data characteristics, and the quantitative objects mainly include ship types, geographic areas and consequence of pirate attack. The output variable calculated by the decision tree method related to consequence of pirate attack is easy to understand, which is used to represent the severity of pirate attack Here, this paper simply states why two factors, ship types and geographic areas, are chosen as the initial data. First reason is that different cargoes are loaded on the types of ships, which enable board height of ships are not same. For instance, the board of ship loading container cargo is higher than the others', because the container weight is lighter. Conversely, the board of ship loading dry and liquid bulk cargo is lower, because this kind of goods is heavier. The advantage of ship with high board is that pirates hardly climb up to embark for hijacking. But the ship with lower board is easy to attempt hijacking. From the perspective of earnings, pirates tend to hijack the ship with lower board generally. Second reason is about the geographic areas. Currently, pirates operate the mother ship mainly, who are going to unionized and scale up, but pirates cannot go to the place where is far away form their base on the coastline generally. That is to say, pirates from the different areas always appear in area where they often go due to the different centralized location of pirates. Furthermore, pirates from the different areas have attack preferences to different ship, which even causes various consequences due to diverse equipment and preferences. In general, ship types and geographic areas have a certain influence on the consequences of pirate attack. In database statistics, there are over 40 statements about ship type; and according to the research results provided by the author, there are 9 movement areas with different characteristics in waters of East Africa. In term of pirate attack consequence, ships are got the different result after robbed by pirates. However, the reports saved in the database describe those situations with some qualitative words, which makes that the two attributes mentioned above cannot be quantified simply. After the data is compared carefully, this paper finally decides to quantify depending on different keywords, and the selected specific keywords are:

1. Is there any sailor died: killed, dead;
2. Is there any sailor injured or missing: injured (badly, seriously), missing.
3. Is there any ship damaged, is there any cargo missing: broken, stolen, miss.

By quantified the basic data, this paper has processed the quantitative outcome further, mainly including categories of ship types and levels of pirate attack consequence. As for categories of ship types, there are over 40 types concluded from the first step, and it shows big quantity variance among each of types. This paper decides to divide these ships according to the type of cargoes after the specific ship type is analyzed. Finally, ships are divided into 5 categories, and the quantity variance of each category has been improved greatly. The specific category result is:

1st category: Ship loading dry and bulk cargo;

2nd category: Ship loading containers;
3rd category: ship loading liquid and bulk cargo;
4th category: ship loading sundry goods;
5th category: the others.

And speak of area selection, this paper divides the waters into numerous disjoint areas. If pirate attacks occur in a certain area, it means the area in attribute is the area number. For example, the waters of East Africa has been divided into 9 valid areas, so that there are 11 possible values from 0 to 11 totally by considering the areas are not classified and out of bounds. At last, the consequence levels are acquired from the keywords statistics, the result are: dead: 11; badly injured: 4; injured: 19; broken: 495; N/A: 670. Considering the casualty status of crew and damaged conditions of ships have different impact, this paper classified the consequences, and the specific levels are: 1st level: no crew casualties and damaged ships (N/A); 2nd level: the crew is injured slightly, or the ship is broken(injured, broken); 3rd level: the crew is injured slightly and the ship is broken (injured, broken); 4th level: Some crew members are dead, or the crew members are injured badly (dead or badly-injured). Based on the levels information, the statistical result is given as below: 1st level: 670 incidents; 2nd level: 486 incidents; 3rd level: 12 incidents; 4th level: 14 incidents.

2.2. Generation and verification of decision tree

The test is operated based on the processed data. By the consequence levels, it is found that the statistical differences between 1st and 2nd levels is large; and relatively speaking, the consequences of 3rd and 4th levels are worse badly. Thus, this paper adopts the sampling method used by re-sample method [15] to duplicate this part of data for enabling the data size of 1st and 2nd levels are equal basically. Afterwards, all the data are processed to generate the decision tree, shown as Fig. 2. The 1st layer is area, and the 2nd layer is ship type.

When the decision tree is constructed, the reasonability of generated decision tree should be verified. In consequences processing of pirates attack, the decision tree is constructed by all the data, this method brings the fitting risk undoubtedly. For this reason, this paper uses [16] the cross validation method to verify the validity of model and algorithm. Cross validation means that the original data is divided into two parts that respective are training set and test set. The training set is used to learn model, and the test set is applied for verifying the model regarded as learning outcome. The measurement standards obtained by this way are more persuasive. The model built by the data A is not used to verify A, but B, so there is no circulatory verification. This paper uses 5-fold cross validation, namely dividing into 5 parts at random. Four parts of them trains one test each time. This step should be repeated 5 times. In this paper, four evaluation criterions are adopted. Accuracy means the correct proportion; precision rate refers to the proportion of true normal sample in the predicted normal samples (precision rate = TP/(TP + FP)); recall rate means the how many normal samples are predicted correctly in all true normal samples (recall rate = TP/(TP + FN)); f1-score is a comprehensive score (fl = (2*precision

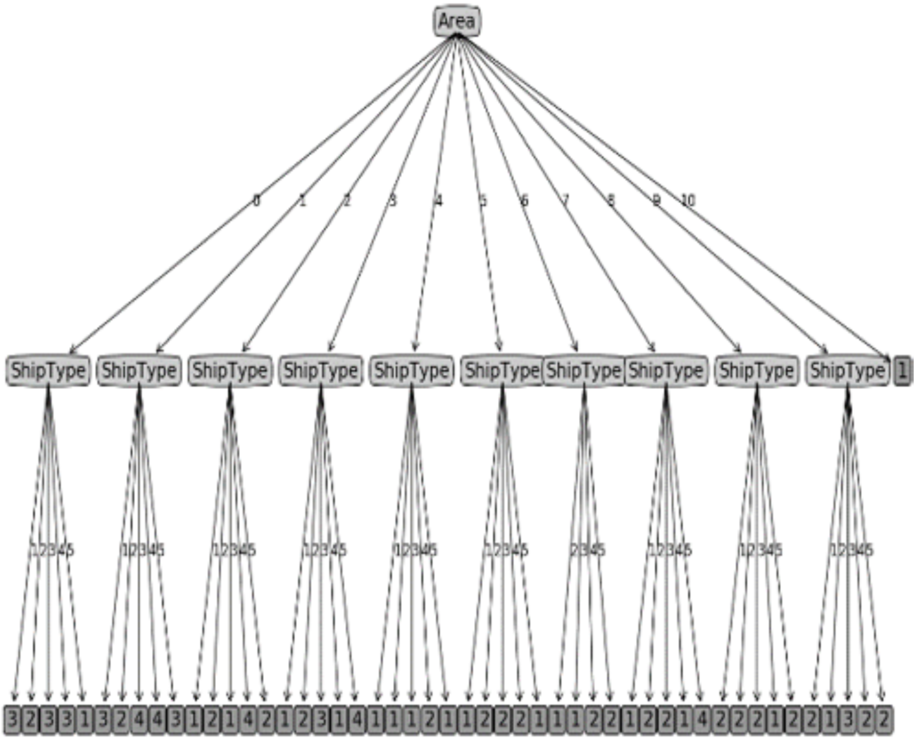


Fig. 2. Decision tree of pirate attack incidents

rate*recall rate/(precision rate+recall rate)). By calculated and verified, it is found that overall comprehensive accuracy is 57.1 %; severity level 1's precision rate, recall rate and f1-score respectively are 49.6 %, 20.2 % and 28.1%; severity level 2's are 55.0 %, 16.4 %, and 23.3 %; severity level 3's are 55.5 %, 84.8 % and 66.9 %; severity level 4's are 63.9 %, 100 % and 77.9 %. Because there are 4 severity levels, precision rate, recall rate and f1-score are only 25 % if predicted at random. It is thus observed that synthetically, the result concluded from the forewarning model of pirate attack with decision tree algorithm is far superior to the results of random guessing. Thus, this model is quite reasonable; especially the precision rate exceeds 50 %.

3. Conclusion

For forewarning transport ships early with the current information before pirates attack, this paper constructs the forewarning model of pirate attack based on the decision tree algorithm. This algorithm mainly considers the types of transport ships and geographical areas, and set them as the input data to predict risk levels of pirate attack. On that basis, this paper executes the practical calculation and verification according the basic data of pirate attack in East Africa area saved in GISIS database.

The verification result shows that compare with the random prediction, this model has obvious advantage because its precision rate of predicting pirate attack is over 50 %. However, due to the data characteristics of East Africa waters, this paper chooses two attributes, ship types and geographic areas, which can improve the predictive effect further, especially in the part of recall rate. The current result has proved the validity of this method. Meanwhile, when this model is used in practice, some diacritical attributes can be added to learn decision tree so as to realize better predictive effect.

References

[1] A. R. Andres, S. A. Asongu: *Fighting software piracy: which governance tools matter in Africa.* Journal of Business Ethics *118* (2013), No. 3, 667–682.

[2] B. S. Sergi, G. Morabito: *The pirates' curse: Economic impacts of the maritime piracy.* Studies in Conflict & Terrorism *39* (2016), No. 10, 935–952.

[3] E. Marchione, S. D. Johnson, A. G. Wilson: *Modelling maritime piracy: A spatial approach.* Journal of Artificial Societies and Social Simulation *17* (2014), No. 2, 1–9.

[4] L. Carral, C. F. Garrido, J. J. de Troya, J. A. Fraguela: *Considering anti-piracy ship security: Citadel design and use.* Brodogradnja/Shipbuilding *66* (2015), No. 3, 75–90.

[5] E. R. Gregorio: *The Filipino seafarers' lived experiences aboard international shipping vessels: A basis for health promotion intervention.* Acta Medica Philippina *45* (2012), No. 3, 69–74.

[6] M. Neocleous: *The universal adversary will attack: pigs, pirates, zombies, Satan and the class war.* Critical Studies on Terrorism *8* (2015), No. 1, 15–32.

[7] J. J. Dabrowski, J. P. de Villiers: *Maritime piracy situation modelling with dynamic Bayesian networks.* International Fusion *23* (2015), 116–130.

[8] J. Wei, F. Wang, M. K Lindell: *The evolution of stakeholders' perceptions of disaster: A model of information flow.* Journal of the Association for Information Science and Technology *67* (2016), No. 2, 441–453.

[9] C. H. Chen, C. M. Hong, T. C. Ou: *Hybrid fuzzy control of wind turbine generator by pitch control using RNN.* International Journal of Ambient Energy *33* (2012), No. 2, 56–64.

[10] S. Ota, Y. Kawatake, R. Kikuchi, K. Tamura: *Analysis on effectiveness of countermeasures against piracy based on incident data.* The Journal of Japan Institute of Navigation *128* (2013), 73–80.

[11] E. D. Carter: *Where's Che? Politics, pop culture, and public memory in Rosario, Argentina.* FOCUS on Geography *55* (2012), No. 1, 1–10.

[12] C. Liu, X. Zhao, M. Li: *Pre-alarm technique for gas disaster and comprehensive solution scheme of computer system construction.* Mining Safety & Environmental Protection *36*(2009), No. s1, 60–63.

[13] L. Nanni, C. Fantozzi, N. Lazzarini: *Coupling different methods for overcoming the class imbalance problem.* Neurocomputing *158* (2015), 48–61.

[14] M. Townsley, A. Oliveira: *Space-time dynamics of maritime piracy.* Security Journal *28* (2015), No. 3, 217–229.

[15] W. Dai, W. Ji: *A mapreduce implementation of C4. 5 decision tree algorithm.* International Journal of Database Theory and Application *7* (2014), No. 1, 49–60.

[16] A. Vehtari, A. Gelman, J. Gabry: *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.* Statistics and Computing *27* (2017), No. 5, 1413–1432.

Received October 12, 2017